

Nonnegative Decomposition of Multivariate Information

Paul L. Williams^{1,*} and Randall D. Beer^{1,2}

¹*Cognitive Science Program and*

²*School of Informatics and Computing*

Indiana University, Bloomington, Indiana 47406 USA

(Dated: April 16, 2010)

Of the various attempts to generalize information theory to multiple variables, the most widely utilized, interaction information, suffers from the problem that it is sometimes negative. Here we reconsider from first principles the general structure of the information that a set of sources provides about a given variable. We begin with a new definition of redundancy as the minimum information that any source provides about each possible outcome of the variable, averaged over all possible outcomes. We then show how this measure of redundancy induces a lattice over sets of sources that clarifies the general structure of multivariate information. Finally, we use this redundancy lattice to propose a definition of partial information atoms that exhaustively decompose the Shannon information in a multivariate system in terms of the redundancy between synergies of subsets of the sources. Unlike interaction information, the atoms of our partial information decomposition are never negative and always support a clear interpretation as informational quantities. Our analysis also demonstrates how the negativity of interaction information can be explained by its confounding of redundancy and synergy.

PACS numbers: 89.70.-a, 87.19.1o, 87.10.Vg, 89.75.-k

Keywords: information theory, interaction information, redundancy, synergy, multivariate interaction

I. INTRODUCTION

From its roots in Shannon’s seminal work on reliability and coding in communication systems, information theory has grown into a ubiquitous general tool for the analysis of complex systems, with application in neuroscience, genetics, physics, machine learning, and many other areas. Somewhat surprisingly, the vast majority of work in information theory concerns only the simplest possible case: the information that a single variable provides about another. This is quantified by Shannon’s mutual information, which is by far the most widely used concept from information theory [1]. The second most popular concept, conditional mutual information, considers interactions between multiple variables in only the most rudimentary sense: it seeks to eliminate the influence of other variables in order to isolate the dependency between two variables of interest. In contrast, many of the most interesting and challenging scientific questions, such as many-body problems in physics [2], n -person games in game theory [3], and population coding in neuroscience [4, 5], involve understanding the structure of interactions between three or more variables.

The two main attempts to generalize information theory to multivariate interactions are the *total correlation* proposed by Watanabe [6] (also known as the multivariate constraint [7], multiinformation [8], and integration [9]) and the *interaction information* of McGill [10] (also known as multiple mutual information [11], co-information [12], and synergy [13]). The total correlation, as its name suggests, measures the total amount of dependency between a set of variables as a single monolithic quantity. Thus, the total correlation does not provide any insight into how dependencies are distributed amongst the variables, i.e., it says nothing about the *structure* of multivariate information.

In contrast, interaction information was proposed as a measure of the amount of information bound up in a set of variables beyond that which is present in any subset of those variables. Thus, entropy and mutual information correspond to first- and second-order interaction information, respectively, and together with its third-, fourth-, and higher-order variants, interaction information provides a way of characterizing the structure of multivariate information. Interaction information is also the natural generalization of mutual information when Shannon entropy is viewed as a signed measure on information diagrams [12, 14, 15]. However, the wider use of interaction information has largely been hampered by the “odd” [12] and “unfortunate” [15] property that, for three or more variables, the interaction information can be negative (see also [11, 14, 16–18]). For information as it is commonly understood, it is entirely unclear what it means for one variable to provide “negative information” about another. Moreover, as we demonstrate below, the confusing property of negativity is actually symptomatic of deeper problems regarding the interpretation of interaction information for larger systems. As a result, there remains no generally accepted extension of information theory for characterizing the structure of multivariate interactions.

Here we formulate a new perspective on the structure of multivariate information. Beginning from first principles, we consider the general structure of the information that a set of sources provide about a given variable. We propose a new definition of redundancy as the minimum information that any source provides about each outcome of the variable, averaged over all possible outcomes. Then we show how this definition can be used to exhaustively decompose the Shannon information in a multivariate system into partial information atoms consisting of redundancies between synergies of subsets of the sources. We also demonstrate that partial information forms a lattice that clarifies the general structure of multivariate information. Unlike interaction information, the atoms of our

* plw@indiana.edu

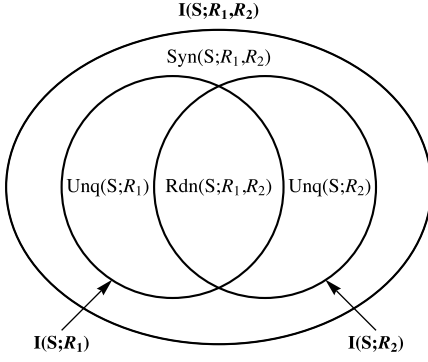


FIG. 1. Structure of multivariate information for 3 variables. Labelled regions correspond to unique information (*Unq*), redundancy (*Rdn*), and synergy (*Syn*).

partial information decomposition are never negative and always support a clear interpretation as informational quantities. Finally, our analysis also demonstrates how the negativity of interaction information can be explained by its confounding of redundant and synergistic interactions.

II. THE STRUCTURE OF MULTIVARIATE INFORMATION

Suppose we are given a random variable S and a random vector $\mathbf{R} = \{R_1, R_2, \dots, R_{n-1}\}$. Then our goal is to decompose the information that \mathbf{R} provides about S in terms of the partial information contributed either individually or jointly by various subsets of \mathbf{R} . For example, in a neuroscience context, S may correspond to a stimulus that takes on different values and \mathbf{R} to the evoked responses of different neurons. In this case, we would like to quantify the information that the joint neural response provides about the stimulus, and to distinguish between information due to responses of individual neurons versus combinations of them [5, 13].

Consider the simplest case of a system with three variables. How much total information does $\mathbf{R} = \{R_1, R_2\}$ provide about S ? How do R_1 and R_2 contribute to the total information? The answer to the first question is given by the mutual information $I(S; R_1, R_2)$, while for the latter we can identify three distinct possibilities. First, R_1 may provide information that R_2 does not, or vice versa (*unique information*). For example, if R_1 is a copy of S and R_2 is a degenerate random variable, then the total information from \mathbf{R} reduces to the unique information from R_1 . Second, R_1 and R_2 may provide the same or overlapping information (*redundancy*). For example, if R_1 and R_2 are both copies of S then they redundantly provide complete information. Third, the combination of R_1 and R_2 may provide information that is not available from either alone (*synergy*). A well-known example for binary variables is the exclusive-OR function $S = R_1 \oplus R_2$, in which case R_1 and R_2 individually provide no information but together provide complete information. Thus, intuitively, the total information from \mathbf{R} decomposes into unique information from R_1 and R_2 , redundant information shared by R_1 and R_2 , and synergistic

information contributed jointly by R_1 and R_2 (FIG. 1).

In sum, for three variables we can identify unique information, redundancy, and synergy as the basic atoms of multivariate information. In fact, as later developments will clarify, unique information is best thought of as a degenerate form of redundancy or synergy, so that redundancy and synergy alone constitute the basic building blocks of multivariate information. In particular, we will find that various combinations of redundancy and synergy, which may at first sound paradoxical, play a fundamental role in structuring multivariate information in higher dimensions. Next we proceed to formalize these ideas, beginning with the problem of defining a measure of redundancy.

III. MEASURING REDUNDANCY

Let $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k$ be nonempty and potentially overlapping subsets of \mathbf{R} , which we call *sources*. How can we quantify the redundant information that all sources provide about S ?

Of course, the information supplied by each \mathbf{A}_i is given simply by $I(S; \mathbf{A}_i)$, the mutual information between S and \mathbf{A}_i . However, it is crucial to note that mutual information is actually a measure of *average* or *expected* information, where the expected value is taken over outcomes of the random variables. Thus, for instance, two sources might provide the same average amount of information, while also providing information about different outcomes of S . Stated formally, the information provided by a source \mathbf{A} can be written as

$$I(S; \mathbf{A}) = \sum_s p(s) I(S = s; \mathbf{A}) \quad (1)$$

where the *specific information* $I(S = s; \mathbf{A})$ quantifies the information associated with a particular outcome s of S . Various definitions of specific information have been proposed to quantify different relationships between S and \mathbf{A} (see Appendix A), but for our purposes the most useful is

$$I(S = s; \mathbf{A}) = \sum_{\mathbf{a}} p(\mathbf{a}|s) \left[\log \frac{1}{p(s)} - \log \frac{1}{p(s|\mathbf{a})} \right]. \quad (2)$$

The term $\frac{1}{p(s)}$ is called the surprise of s , so $I(S = s; \mathbf{A})$ is the average reduction in surprise of s given knowledge of \mathbf{A} . In other words, $I(S = s; \mathbf{A})$ quantifies the information that \mathbf{A} provides about each particular outcome $s \in S$, while $I(S; \mathbf{A})$ is the expected value of this quantity over all outcomes of S .

Given these considerations, a natural measure of redundancy is the expected value of the minimum information that any source provides about each outcome of S , or

$$I_{\min}(S; \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k\}) = \sum_s p(s) \min_{\mathbf{A}_i} I(S = s; \mathbf{A}_i). \quad (3)$$

I_{\min} captures the idea that redundancy is the information common to all sources (the minimum information that any source provides), while taking into account that sources may provide information about different outcomes of S . Note that, like the

mutual information, I_{\min} is also an expected value of specific information terms.

I_{\min} also has several important properties that further support its interpretation as a measure of redundancy. First, I_{\min} is nonnegative, a property that follows directly from the non-negativity of specific information (see Appendix D). Second, I_{\min} is less than or equal to $I(S; \mathbf{A}_i)$ for all \mathbf{A}_i 's, with equality if and only if $I(S = s; \mathbf{A}_i) = I(S = s; \mathbf{A}_j)$ for all i and j and all $s \in S$. Thus, as one would hope, the amount of redundant information is bounded by the information provided by each source, with equality if and only if all sources provide the exact same information about S . Finally, and closely related to the previous property, for a given source \mathbf{A} the amount of information redundant with \mathbf{A} is maximal for $I_{\min}(S; \{\mathbf{A}\}) = I(S; \mathbf{A})$. In other words, redundant information is maximized by the “self-redundancy,” analogous to the property that mutual information is maximized by the self-information $I(S; S) = H(S)$.

What are the distinct ways in which collections of sources might contribute redundant information? Formally, answering this question means identifying the domain of I_{\min} . Thus far, we have assumed that the natural domain is the collection of all possible sets of sources, but in fact this can be greatly simplified. To illustrate, consider two sources, \mathbf{A} and \mathbf{B} , with \mathbf{A} a subset of \mathbf{B} . Clearly, any information provided by \mathbf{A} is also provided by \mathbf{B} , so the redundancy between \mathbf{A} and \mathbf{B} reduces to the self-redundancy for \mathbf{A} ,

$$I_{\min}(S; \{\mathbf{A}, \mathbf{B}\}) = I_{\min}(S; \{\mathbf{A}\}) = I(S; \mathbf{A}).$$

Furthermore, for any source \mathbf{C} , it follows that $I_{\min}(S; \{\mathbf{A}, \mathbf{B}, \mathbf{C}\}) = I_{\min}(S; \{\mathbf{A}, \mathbf{C}\})$. Extending this idea, for any collection of sources where some are supersets of others, the redundancy for that collection is equivalent to the redundancy with all supersets removed. Thus, the domain for I_{\min} can be reduced to the collection of all sets of sources such that no source is a superset of any other. Formally, this set can be written as

$$\mathcal{A}(\mathbf{R}) = \{\alpha \in \mathcal{P}_1(\mathcal{P}_1(\mathbf{R})) : \forall \mathbf{A}_i, \mathbf{A}_j \in \alpha, \mathbf{A}_i \not\subseteq \mathbf{A}_j\}, \quad (4)$$

where $\mathcal{P}_1(\mathbf{R}) = \mathcal{P}(\mathbf{R}) \setminus \{\emptyset\}$ is the set of all nonempty subsets of \mathbf{R} . Henceforth, we will denote elements of $\mathcal{A}(\mathbf{R})$, corresponding to collections of sources, with bracketed expressions containing only the indices for each source. For instance, $\{\{R_1, R_2\}\}$ will be $\{12\}$, $\{\{R_1\}, \{R_2, R_3\}\}$ will be $\{1\}\{23\}$, and so forth.

The possibilities for redundancy are also naturally structured, which is shown by extending the same line of reasoning to define an ordering \preceq on the elements of $\mathcal{A}(\mathbf{R})$. Consider two collections of sources, $\alpha, \beta \in \mathcal{A}(\mathbf{R})$, where for each source $\mathbf{B} \in \beta$ there exists a source $\mathbf{A} \in \alpha$ with \mathbf{A} a subset of \mathbf{B} . This means that for each source $\mathbf{B} \in \beta$ there is a source $\mathbf{A} \in \alpha$ such that \mathbf{A} provides no more information than \mathbf{B} . The redundant information shared by all $\mathbf{B} \in \beta$ must therefore at least include any redundant information shared by all $\mathbf{A} \in \alpha$. Thus, we can define a partial order over the elements of $\mathcal{A}(\mathbf{R})$ such that one element (collection of sources) is considered to precede another if and only if the latter provides any redundant information

that the former provides. The ordering relation \preceq is formally defined as

$$\forall \alpha, \beta \in \mathcal{A}(\mathbf{R}), (\alpha \preceq \beta \Leftrightarrow \forall \mathbf{B} \in \beta, \exists \mathbf{A} \in \alpha, \mathbf{A} \subseteq \mathbf{B}). \quad (5)$$

Applying this ordering to the elements of $\mathcal{A}(\mathbf{R})$ produces a *redundancy lattice*, in which a higher element provides at least as much redundant information as a lower one (FIG. 2; see Appendix C).

The redundancy lattice provides a wealth of insight into the structure of redundancy. For instance, from the redundancy lattice it is possible to read off some of the properties of I_{\min} noted earlier. The property that redundancy for a source is maximized by the self-redundancy can be seen from the fact that any node corresponding to an individual source appears higher in the redundancy lattice than any other node involving that source. For example, in FIG. 2B, the node labeled $\{12\}$, corresponding to the self-redundancy for the source $\{R_1, R_2\}$, occurs higher than nodes labeled $\{12\}\{13\}$, $\{12\}\{13\}\{23\}$, and $\{3\}\{12\}$. Another property of I_{\min} that can be seen from these diagrams relates to the top and bottom elements of the lattice. The top element corresponds to the self-redundancy for \mathbf{R} , reflecting the fact that I_{\min} is bounded from above by the total amount of information provided by \mathbf{R} . At the other end of the spectrum, the bottom element corresponds to the redundant information that each individual element of \mathbf{R} provides, with all other possibilities for redundancy falling between these two extremes.

IV. PARTIAL INFORMATION DECOMPOSITION

The redundant information associated with each node of the redundancy lattice includes, but is not limited to, the redundant information provided by all nodes lower in the lattice. Thus, moving from node to node up the lattice, I_{\min} can be thought of as a kind of “cumulative information function,” effectively integrating the information provided by increasingly inclusive collections of sources. Next, we derive an inverse of I_{\min} called the partial information function (PI-function). Whereas I_{\min} quantifies cumulative information, the PI-function measures the partial information contributed uniquely by each particular collection of sources. This partial information will form the atoms into which we decompose the total information that \mathbf{R} provides about S .

For a collection of sources $\alpha \in \mathcal{A}(\mathbf{R})$, the PI-function, denoted $\Pi_{\mathbf{R}}$, is defined implicitly by

$$I_{\min}(S; \alpha) = \sum_{\beta \preceq \alpha} \Pi_{\mathbf{R}}(S; \beta). \quad (6)$$

Formally, $\Pi_{\mathbf{R}}$ corresponds to the Möbius inverse of I_{\min} [19, 20]. From this relationship, it is clear that $\Pi_{\mathbf{R}}$ can be calculated recursively as

$$\Pi_{\mathbf{R}}(S; \alpha) = I_{\min}(S; \alpha) - \sum_{\beta \prec \alpha} \Pi_{\mathbf{R}}(S; \beta). \quad (7)$$

Put into words, $\Pi_{\mathbf{R}}(S; \alpha)$ quantifies the information provided redundantly by the sources of α that is not provided by any

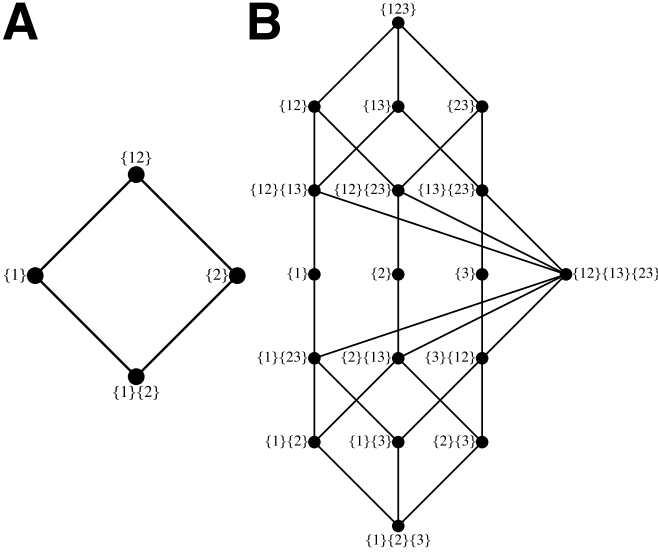


FIG. 2. Redundancy lattice for (A) 3 and (B) 4 variables.

simpler collection of sources (i.e., any β lower than α on the redundancy lattice). In Appendix D, it is shown that $\Pi_{\mathbf{R}}$ can be written in closed form as

$$\Pi_{\mathbf{R}}(S; \alpha) = I_{\min}(S; \alpha) - \sum_s p(s) \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}) \quad (8)$$

where α^- represents the nodes immediately below α in the redundancy lattice. From this formulation, it is readily shown that $\Pi_{\mathbf{R}}$ is nonnegative (see Appendix D), and thus can be naturally interpreted as an informational quantity associated with the sources of α .

The decomposition of mutual information into a sum of PI-terms follows from

$$I(S; \mathbf{A}) = I_{\min}(S; \{\mathbf{A}\}) = \sum_{\beta \preceq \{\mathbf{A}\}} \Pi_{\mathbf{R}}(S; \beta). \quad (9)$$

For the 3-variable case $\mathbf{R} = \{R_1, R_2\}$, Equation (9) yields

$$I(S; R_1) = \Pi_{\mathbf{R}}(S; \{1\}) + \Pi_{\mathbf{R}}(S; \{1\}\{2\}) \quad (10)$$

and

$$I(S; R_1, R_2) = \Pi_{\mathbf{R}}(S; \{1\}) + \Pi_{\mathbf{R}}(S; \{2\}) + \Pi_{\mathbf{R}}(S; \{1\}\{2\}) + \Pi_{\mathbf{R}}(S; \{12\}). \quad (11)$$

The relationship between these equations can be represented as a *partial information (PI) diagram* (FIG. 3A), which illustrates the way in which the total information that \mathbf{R} provides about S is distributed across various combinations of sources. Furthermore, comparing this diagram with FIG. 1 makes immediately clear the meaning of each partial information term. First, from Equation (8), we have that $\Pi_{\mathbf{R}}(S; \{1\}\{2\}) = I_{\min}(S; \{1\}\{2\})$, which, from the definition of I_{\min} , corresponds to the redundancy for R_1 and R_2 . The unique information for R_1 is given by $\Pi_{\mathbf{R}}(S; \{1\}) = I(S; R_1) - I_{\min}(S; \{1\}\{2\})$, which is the total information from R_1 minus the redundancy, and likewise for R_2 . Finally,

the additional information provided by the combination of R_1 and R_2 is given by $\Pi_{\mathbf{R}}(S; \{12\})$, corresponding to their synergy.

To fix ideas, consider the example in FIG. 4A. From the symmetry of the distribution, it is clear that R_1 and R_2 must provide the same amount of information about S . Indeed, this is easily verified, with $I(S; R_1) = I(S; R_2) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3}$. However, it is also clear that R_1 and R_2 provide information about different outcomes of S . In particular, given knowledge of R_1 , one can determine conclusively whether or not outcome $S = 2$ occurs (which is not the case for R_2), and likewise for R_2 and outcome $S = 1$. This feature is captured by $\Pi_{\mathbf{R}}(S; \{1\}) = \Pi_{\mathbf{R}}(S; \{2\}) = \frac{1}{3}$, indicating that R_1 and R_2 each provide $\frac{1}{3}$ bits of unique information about S . The redundant information, $\Pi_{\mathbf{R}}(S; \{1\}\{2\}) = \log 3 - \log 2$, captures the fact that knowledge of either R_1 or R_2 reduces uncertainty about S from three equally likely outcomes to two. Finally, R_1 and R_2 also provide $\frac{1}{3}$ bits of synergistic information, i.e., $\Pi_{\mathbf{R}}(S; \{12\}) = \frac{1}{3}$. This value reflects the fact that R_1 and R_2 together uniquely determine whether or not outcome $S = 0$ occurs, which is not true for R_1 or R_2 alone.

Note that, unlike mutual information or interaction information, partial information is *not* symmetric. For instance, the synergistic information that R_1 and R_2 provide about S is not in general equal to the synergistic information that S and R_2 provide about R_1 . This property is also illustrated by the example in FIG. 4A. Given knowledge of S , one can uniquely determine the outcome of R_1 (and R_2), so that S provides complete information about both. Thus, it is not possible for the combination of S and R_2 to provide any additional synergistic information about R_1 , since there is no remaining uncertainty about R_1 when S is known. In contrast, as was just noted, R_1 and R_2 provide $\frac{1}{3}$ bits of synergistic information about S . This asymmetry accounts for our decision to focus on information *about* a particular variable S throughout, since in general the analysis will differ depending on the variable of interest. Note that total information is also asymmetric in this sense, i.e., in general $I(S; R_1, R_2) \neq I(R_1; S, R_2)$ (though, of course, it is symmetric in the sense that $I(S; R_1, R_2) = I(R_1, R_2; S)$).

The general structure of PI-diagrams becomes clear when we consider the decomposition for four variables (FIG. 3B). First, note that all of the possibilities for three variables are again present for four. In particular, each element of \mathbf{R} can provide unique information (regions labeled $\{1\}$, $\{2\}$, and $\{3\}$), information redundantly with one other variable ($\{1\}\{2\}$, $\{1\}\{3\}$, and $\{2\}\{3\}$), or information synergistically with one other variable ($\{12\}$, $\{13\}$, and $\{23\}$). Additionally, information can be provided redundantly by all three variables ($\{1\}\{2\}\{3\}$) or provided by their three-way synergy ($\{123\}$). More interesting are the new kinds of terms representing combinations of redundancy and synergy. For instance, the regions marked $\{1\}\{23\}$, $\{2\}\{13\}$, and $\{3\}\{12\}$ represent information that is available redundantly from either one variable considered individually or the other two considered together. Or, for instance, the region labeled $\{12\}\{13\}\{23\}$ represents the information provided redundantly by the three possible two-way synergies. In general, the PI-atom for a collection of sources corresponds

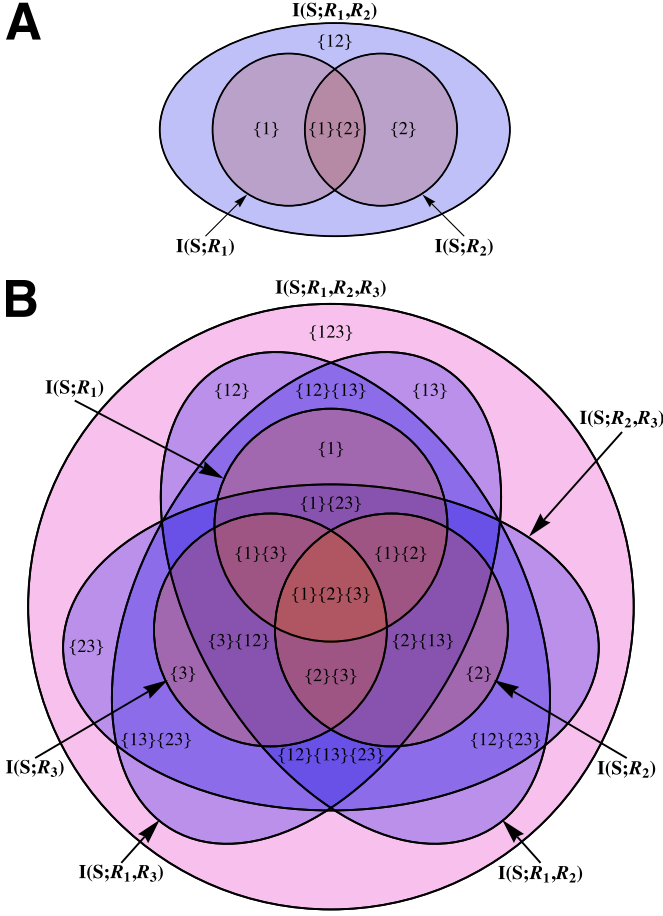


FIG. 3. Partial information diagrams for (A) 3 and (B) 4 variables.

to the information provided redundantly by the synergies of all sources in the collection. This point also clarifies our earlier claim that unique information is best thought of as a degenerate case: unique information corresponds to the combination of first-order redundancy and first-order synergy.

In general, a PI-diagram for n variables, S and $\mathbf{R} = \{R_1, R_2, \dots, R_{n-1}\}$, consists of the following (see Fig. S2 in Appendix E). First, for each element $R_i \in \mathbf{R}$ there is a region corresponding to $I(S; R_i)$. Then, for every subset \mathbf{A} of \mathbf{R} with two or more elements, $I(S; \mathbf{A})$ is depicted as a region containing $I(S; A)$ for all $A \in \mathbf{A}$ but not coextensive with $\bigcup_{A \in \mathbf{A}} I(S; A)$. The difference between $I(S; \mathbf{A})$ and $\bigcup_{A \in \mathbf{A}} I(S; A)$ represents the synergy for \mathbf{A} , the information gained from the combined knowledge of all elements in \mathbf{A} that is not available from any subset. In addition, regions of the diagram intersect generically, representing all possibilities for redundancy. In total, a PI-diagram is composed of the $(n-1)$ -th Dedekind number [21] of PI-atoms, same as the cardinality of $\mathcal{A}(\mathbf{R})$ (see Appendix C). As described above, each PI-atom represents the redundancy of synergies for a particular collection of sources, corresponding to one distinct way for the components of \mathbf{R} to contribute information about S .

Finally, it is instructive to consider the relationship between the redundancy lattice and PI-diagram for n variables.

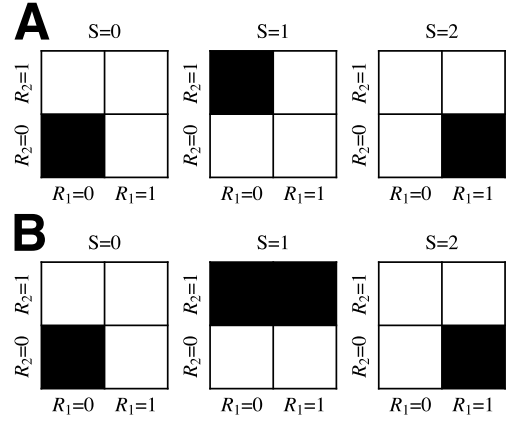


FIG. 4. Probability distributions for $S \in \{0, 1, 2\}$ and $R_1, R_2 \in \{0, 1\}$. Black tiles represent equiprobable outcomes. White tiles are zero-probability outcomes.

First, we note that I_{\min} is analogous to set intersection for PI-diagrams, consistent with the idea of redundancy as overlapping information. Specifically, $I_{\min}(S; \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k\})$ corresponds to the region $\bigcap_i I(S; \mathbf{A}_i)$. From this correspondence between I_{\min} and set intersection, we can establish the following connection: for $\alpha, \beta \in \mathcal{A}(\mathbf{R})$, α is lower than β in the redundancy lattice if and only if $\bigcap_{\mathbf{A} \in \alpha} I(S; \mathbf{A})$ is a subset of $\bigcap_{\mathbf{B} \in \beta} I(S; \mathbf{B})$ in the PI-diagram. Consequently, the redundancy lattice and PI-diagram can be viewed as complementary representations of the same structure, with the PI-diagram a collapsed version of the redundancy lattice formed by embedding regions according to the lattice ordering.

V. WHY INTERACTION INFORMATION IS SOMETIMES NEGATIVE

We next show how PI-decomposition can be used to understand the conditions under which interaction information, the standard generalization of mutual information to multivariate interactions, is negative. The interaction information for three variables is given by

$$I(S; R_1; R_2) = I(S; R_1|R_2) - I(S; R_1) \quad (12)$$

and for $n > 3$ variables is defined recursively as

$$I(S; R_1; R_2; \dots; R_{n-1}) = I(S; R_1; R_2; \dots; R_{n-2}|R_{n-1}) - I(S; R_1; R_2; \dots; R_{n-2}) \quad (13)$$

where the conditional interaction information is defined by simply including the conditioning in all terms of the original definition. Interaction information is symmetric for all permutations of its arguments, and is traditionally interpreted as the information shared by all n variables beyond that which is shared by any subset of those variables.

For 3-variable interaction information, a positive value is naturally interpreted as indicating a situation in which any one variable of the system enhances the correlation between the

other two. For example, a positive value for Equation (12) indicates that knowledge of R_2 enhances the correlation between S and R_1 (and likewise for all other variable permutations). Thus, in the terminology used here, a positive value for $I(S; R_1; R_2)$ indicates the presence of synergy. On the other hand, a negative value for $I(S; R_1; R_2)$ indicates a situation in which any one variable accounts for or “explains away” [22] the correlation between the other two. In other words, a negative value for $I(S; R_1; R_2)$ indicates redundancy. Indeed, $I(S; R_1; R_2)$ is a widely used measure of synergy in neuroscience, where it is interpreted in exactly this way [23–26].

The PI-decomposition for 3-variable interaction information (FIG. 5A; see also Fig. S3 in Appendix E) confirms this interpretation, with $I(S; R_1; R_2)$ equal to the difference between the synergistic and the redundant information, i.e.,

$$I(S; R_1; R_2) = \Pi_{\mathbf{R}}(S; \{12\}) - \Pi_{\mathbf{R}}(S; \{1\}\{2\}). \quad (14)$$

Thus, it is indeed the case that positive values indicate synergy and negative values indicate redundancy.

However, PI-decomposition also makes clear that $I(S; R_1; R_2)$ confounds redundancy and synergy, with the meaning of interaction information ambiguous for any system that exhibits a mixture of the two (cf. [27], who suggest the possibility of mixed redundancy and synergy, but without attempting to disentangle them). For instance, consider again the example in FIG. 4A. As described earlier, in this case R_1 and R_2 provide $\log 3 - \log 2$ bits of redundant information and $\frac{1}{3}$ bits of synergistic information. Consequently, $I(S; R_1; R_2)$ is negative because there is more redundancy than synergy, despite the fact that the system clearly exhibits synergistic interactions. As a second example, consider the distribution in FIG. 4B. In this case, R_1 and R_2 provide $\frac{1}{2}$ bits of redundant information, corresponding to the fact that knowledge of either R_1 or R_2 reduces uncertainty about the outcomes $S = 0$ and $S = 2$. Additionally, R_1 and R_2 provide $\frac{1}{2}$ bits of synergistic information, reflecting the fact that R_1 and R_2 together provide complete information about outcomes $S = 0$ and $S = 2$, which is not true for either alone. Thus, the interaction information in this case is equal to zero despite the presence of both redundant and synergistic interactions, because redundancy and synergy are balanced.

The situation is worse for four-variable interaction information, which is known to violate the interpretation that positive values indicate (pure) synergy and negative values indicate (pure) redundancy [12, 28]. To demonstrate, consider the case of 3-parity, which is the higher-order form of the exclusive-OR, or 2-parity, function mentioned earlier. In this case, we have a system of four binary random variables, S and $\mathbf{R} = \{R_1, R_2, R_3\}$, where the eight outcomes for \mathbf{R} are equiprobable and $S = R_1 \oplus R_2 \oplus R_3$. Intuitively, this corresponds to a case of pure synergy, since the value of S can be determined only when all of the R_i are known. Indeed, using Eq. (13) we find that $I(S; R_1; R_2; R_3)$ for this system is equal to +1 bit, as expected from the interpretation that positive values indicate synergy. However, now consider a second system of binary variables, this time where the two outcomes of S are equiprobable and R_1, R_2 , and R_3 are all copies of S . Clearly this corresponds to a case of pure redun-

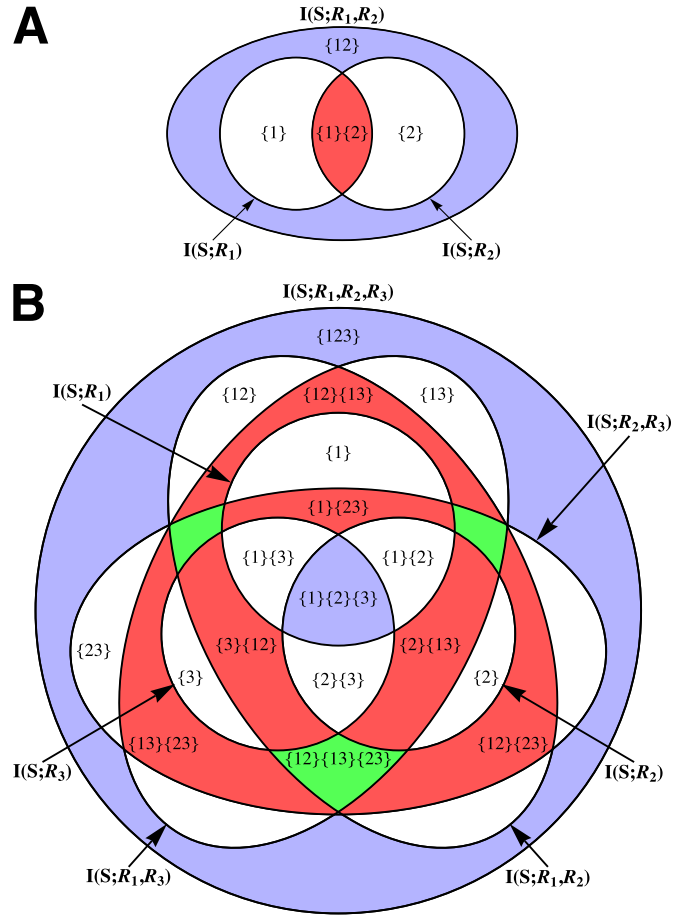


FIG. 5. PI-decomposition of interaction information for (A) 3 and (B) 4 variables. Blue and red regions represent PI-terms that are added and subtracted, respectively. The green region in (B) represents a PI-term that is subtracted twice.

dancy, since the value of S can be determined uniquely from knowledge of any R_i , but $I(S; R_1; R_2; R_3)$ for this system is again equal to +1 bit, same as the case of pure synergy. Thus, a completely redundant system is assigned a positive value for the interaction information, in clear violation of the idea that redundancy is indicated by negative values. Worse still, the 4-variable interaction information fails to distinguish between the polar opposites of purely synergistic and purely redundant information.

The PI-decomposition for 4-variable interaction information (FIG. 5B; see also Fig. S4 in Appendix E) clarifies why this is the case. In terms of PI-atoms, $I(S; R_1; R_2; R_3)$ is given by

$$\begin{aligned} & \Pi_{\mathbf{R}}(S; \{123\}) + \Pi_{\mathbf{R}}(S; \{1\}\{2\}\{3\}) \\ & - \Pi_{\mathbf{R}}(S; \{1\}\{23\}) - \Pi_{\mathbf{R}}(S; \{2\}\{13\}) - \Pi_{\mathbf{R}}(S; \{3\}\{12\}) \\ & - \Pi_{\mathbf{R}}(S; \{12\}\{13\}) - \Pi_{\mathbf{R}}(S; \{12\}\{23\}) - \Pi_{\mathbf{R}}(S; \{13\}\{23\}) \\ & - 2 \times \Pi_{\mathbf{R}}(S; \{12\}\{13\}\{23\}). \end{aligned} \quad (15)$$

Thus, $I(S; R_1; R_2; R_3)$ is equal to the sum of third-order synergy ($\{123\}$) and third-order redundancy ($\{1\}\{2\}\{3\}$), minus the information provided redundantly by a first- and second-order synergy ($\{1\}\{23\}$, $\{2\}\{13\}$, and $\{3\}\{12\}$), minus the

information provided redundantly by two second-order synergies ($\{12\}\{13\}$, $\{12\}\{23\}$, and $\{13\}\{23\}$), and minus twice the information provided redundantly by all three second-order synergies ($\{12\}\{13\}\{23\}$). Thus, systems with pure synergy and pure redundancy have the same value for $I(S; R_1; R_2; R_3)$ because 4-variable interaction information adds in the highest-order synergy and redundancy terms. More generally, the PI-decomposition for $I(S; R_1; R_2; R_3)$ shows why it is difficult to interpret as a meaningful quantity, and as one might expect the story only becomes more complicated in higher dimensions. Thus, although one can readily decompose interaction information into a collection of partial information contributions, and understand the conditions under which it will be positive or negative depending on the relative magnitudes of these contributions, the utility of interaction information for larger systems is unclear.

VI. DISCUSSION

The main objective of this paper has been to quantify multivariate information in such a way that the structure of variable interactions is illuminated. This was accomplished by first defining a general measure of redundant information, I_{\min} , which satisfies a number of intuitive properties for a measure of redundancy. Next, it was shown that I_{\min} induces a lattice structure over the set of possible information sources, referred to as the redundancy lattice, which characterizes the distinct ways that information can be distributed amongst a set of sources. From this lattice, a measure of partial information was derived that captures the unique information contributed by each possible combination of sources. It was then shown that mutual information decomposes into a sum of these partial information terms, so that the total information provided by a source is broken down into a collection of partial information contributions. Moreover, it was demonstrated that each of these terms supports clear interpretation as a particular combination of redundant and synergistic interactions between specific subsets of variables. Finally, we discussed the relationship between partial information decomposition and interaction information, the current de facto measure of multivariate interactions, and used partial information to clarify the confusing property that interaction information is sometimes negative.

One obvious challenge with applying these ideas is that the number of partial information terms grows rapidly for larger

systems. For instance, with 9 variables there are more than 5×10^{22} possibilities [29], and beyond that the Dedekind numbers are not even currently known. Thus, clearly an important direction for future work is to determine efficient ways of calculating partial information terms for larger systems. To this end, the lattice structure of the terms is likely to play an essential role. As with any ordered data structure, the fact that the space of possibilities is highly organized can be readily exploited for efficient use. For instance, as a simple example, if I_{\min} is calculated in a descending fashion over the nodes of the redundancy lattice and at a certain juncture has a value of zero, all of the terms below that node can immediately be eliminated simply from the monotonicity of I_{\min} (see Appendix D). Moreover, if the Markov property or any other constraints hold between the variables, many of the possible partial information terms can also be excluded. Finally, these considerations notwithstanding, it should also be emphasized that 3-variable interaction is the current state of the art, and thus even the simplest form of partial information decomposition can be used to address a number of outstanding questions.

In physics, for example, 3-variable interactions have been explored in relation to the non-separability of quantum systems [30] and in the study of many-body correlation effects [31]. In neuroscience, the concepts of synergy and redundancy for three variables have been examined in the context of neural coding in a number of theoretical and empirical investigations [23–26, 32, 33]. In genetics, multivariate dependencies arise in the analysis of gene-gene and gene-environment interactions in studies of human disease susceptibility [28, 34, 35]. Moreover, similar issues have also been explored in machine learning [22, 27, 36], ecology [37], quantum information theory [38], information geometry [39], rough set analysis [40], and cooperative game theory [41]. Thus, in all of these cases, the 3-variable form of partial information decomposition can be applied immediately to illuminate the structure of multivariate dependencies, while the general form provides a clear way forward in the study of more complex systems of interactions.

ACKNOWLEDGMENTS

We thank O. Sporns, J. Beggs, A. Kolchinsky, and L. Yaeger for helpful comments. This work was supported in part by NSF grant IIS-0916409 (to R.D.B.) and an NSF IGERT traineeship (to P.L.W.).

-
- [1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (Univ of Illinois Press, 1949).
 - [2] D. Pines, *The Many-Body Problem* (Addison-Wesley, 1997).
 - [3] R. D. Luce and H. Raiffa, *Games and Decisions: Introduction and Critical Survey* (Dover, 1989).
 - [4] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, 2001).
 - [5] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, 1999).
 - [6] S. Watanabe, IBM Journal of Research and Development, **4**, 66 (1960).
 - [7] W. R. Garner, *Uncertainty and Structure as Psychological Concepts* (Wiley, 1962).
 - [8] M. Studeny and J. Vejnarova, Learning in Graphical Models, 261 (1998).
 - [9] G. Tononi, O. Sporns, and G. M. Edelman, Proc Natl Acad Sci USA, **91**, 5033 (1994).
 - [10] W. J. McGill, Psychometrika, **19**, 97 (1954).

- [11] T. S. Han, *Information and Control*, **46**, 26 (1980).
- [12] A. J. Bell, *Proceedings of ICA2003*, 921 (2003).
- [13] T. J. Gawne and B. J. Richmond, *J Neurosci*, **13**, 2758 (1993).
- [14] R. W. Yeung, *Information Theory and Network Coding* (Springer, 2008).
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, 2006).
- [16] S. Takano, *Proc Jpn Acad*, **50**, 109 (1974).
- [17] T. Tsujishita, *Adv Appl Math*, **16**, 269 (1995).
- [18] Z. Zhang and R. W. Yeung, *IEEE Trans Inf Theory*, **44**, 1440 (1998).
- [19] G. Rota, *Probability Theory and Related Fields*, **2**, 340 (1964).
- [20] R. P. Stanley, *Enumerative Combinatorics*, Vol. 1 (Cambridge Univ Press, 1997).
- [21] L. Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions* (Springer, 1974).
- [22] J. Pearl and G. Shafer, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 1988).
- [23] N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. de Ruyter van Steveninck, *Neural Comput*, **12**, 1531 (2000).
- [24] S. Panzeri, S. R. Schultz, A. Treves, and E. T. Rolls, *Proc R Soc B*, **266**, 1001 (1999).
- [25] E. Schneidman, W. Bialek, and M. J. Berry, *J Neurosci*, **23**, 11539 (2003).
- [26] P. E. Latham and S. Nirenberg, *J Neurosci*, **25**, 5195 (2005).
- [27] A. Jakulin and I. Bratko, *Arxiv preprint cs/0308002* (2003).
- [28] D. Anastassiou, *Mol Syst Biol*, **3**, 1 (2007).
- [29] D. Wiedemann, *Order*, **8**, 5 (1991).
- [30] N. J. Cerf and C. Adami, *Phys Rev A*, **55**, 3371 (1997).
- [31] H. Matsuda, *Phys Rev E*, **62**, 3096 (2000).
- [32] I. Gat and N. Tishby, *Advances in NIPS*, 111 (1999).
- [33] N. S. Narayanan, E. Y. Kimchi, and M. Laubach, *J Neurosci*, **25**, 4207 (2005).
- [34] J. H. Moore, J. C. Gilbert, C. T. Tsai, F. T. Chiang, T. Holden, N. Barney, and B. C. White, *J Theor Biol*, **241**, 252 (2006).
- [35] P. Chanda, A. Zhang, D. Brazeau, L. Sucheston, J. L. Freudenheim, C. Ambrosone, and M. Ramanathan, *Am J Hum Genet*, **81**, 939 (2007).
- [36] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge Univ Press, 2003).
- [37] L. Orlóci, M. Anand, and V. D. Pillar, *Community Ecol*, **3**, 217 (2002).
- [38] V. Vedral, *Rev Mod Phys*, **74**, 197 (2002).
- [39] S. Amari, *IEEE Trans Inf Theory*, **47**, 1701 (2001).
- [40] G. Gediga and I. Düntsch, in *Rough-Neural Computing*, edited by S. K. Pal, L. Polkowski, and A. Skowron (Physica Verlag, Heidelberg, 2003).
- [41] M. Grabisch and M. Roubens, *Int J Game Theory*, **28**, 547 (1999).
- [42] M. R. DeWeese and M. Meister, *Network*, **10**, 325 (1999).
- [43] D. A. Butts, *Network*, **14**, 177 (2003).
- [44] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, 2nd ed. (Cambridge Univ Press, 2002).
- [45] G. A. Grätzer, *General Lattice Theory*, 2nd ed. (Birkhäuser, 2003).
- [46] J. Crampton and G. Loizou, "Two partial orders on the set of antichains," (2000), research note.
- [47] J. Crampton and G. Loizou, *International Mathematical Journal*, **1**, 223 (2001).
- [48] S. M. Ross, *A First Course in Probability*, 8th ed. (Prentice Hall, 2009).

Appendix A: Measures of Specific Information

Measures of specific information are discussed in [42] in the context of quantifying the information that specific neural responses provide about a stimulus ensemble. For random variables S and R , representing stimuli and responses, respectively, the information that R provides about S is decomposed according to

$$I(S; R) = \sum_{r \in R} p(r) i_r(r) \quad (A1)$$

and

$$i_r(r) = H(S) - H(S|r) \quad (A2)$$

where $H(S)$ is the entropy of S and $i_r(r)$ is the *response-specific information* associated with each $r \in R$. The response-specific information quantifies the change in uncertainty about S when response r is observed. In [42], it is shown that i_r is the unique measure of specific information that satisfies additivity, though it is also possible for i_r to be negative.

To distinguish the different role played by stimuli as opposed to responses, an alternative measure of specific information is proposed in [43]. The *stimulus-specific information* for an outcome $s \in S$ is defined as

$$i_s(s) = \sum_{r \in R} p(r|s) i_r(r). \quad (A3)$$

Like the response-specific information, the weighted average of $i_s(s)$ gives the mutual information $I(S; R)$. Stimulus-specific information quantifies the extent to which a particular stimulus s tends to evoke responses that are informative about the entire ensemble S (responses with high values for i_r).

Finally, both [42] and [43] also discuss $I(S = s; R)$, the measure of specific information used here (Eq. (2)). In [43], $I(S = s; R)$ is described as the reduction in surprise of a particular stimulus s gained from each response, averaged over all responses associated with that stimulus. Thus, whereas $i_s(s)$ weights each response r according to the information that it contributes about the entire ensemble S , $I(S = s; R)$ quantifies only the information that R provides about the particular outcome $S = s$. In [42], it is proven that $I(S = s; R)$ is the only measure of specific information that is strictly nonnegative.

Appendix B: Lattice Theory Definitions

Here we review only the basic concepts of lattice theory needed for supporting proofs. For a thorough treatment, see [44, 45].

Definition 1. A pair $\langle X, \leq \rangle$ is a partially ordered set or poset if \leq is a binary relation on X that is reflexive, transitive and antisymmetric.

Definition 2. Let $Y \subseteq X$. Then $a \in Y$ is a maximal element in Y if for all $b \in Y$, $a \leq b \Rightarrow a = b$. A minimal element is defined dually. We denote the set of maximal elements of Y by \bar{Y} and the set of minimal elements by \underline{Y} .

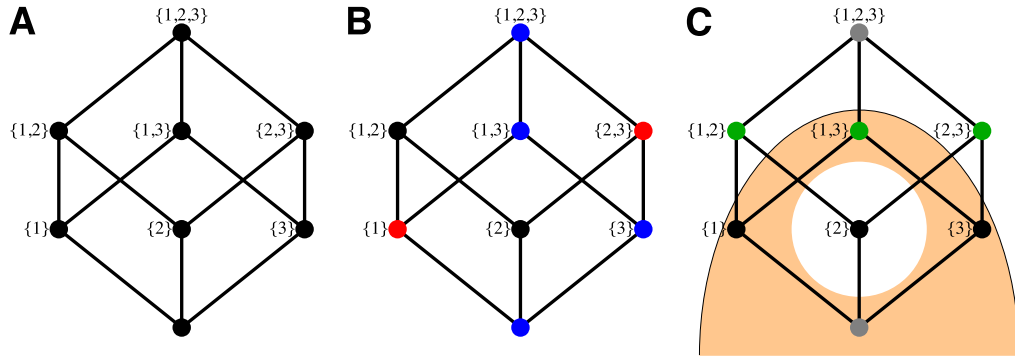


FIG. S1: Basic lattice-theoretic concepts. (A) Hasse diagram of the lattice $\langle \mathcal{P}(X), \subseteq \rangle$ for $X = \{1, 2, 3\}$. (B) An example of a chain (blue nodes) and an antichain (red nodes). (C) The top \top and bottom \perp are shown in gray. Green nodes correspond to $\{1, 2, 3\}^-$, the set of elements covered by $\{1, 2, 3\}$. The orange region represents $\downarrow \{1, 3\}$, the down-set of $\{1, 3\}$.

Definition 3. Let $\langle X, \leq \rangle$ be a poset, and let $Y \subseteq X$. An element $x \in X$ is an upper bound for Y if for all $y \in Y, y \leq x$. A lower bound for Y is defined dually.

Definition 4. An element $x \in X$ is the least upper bound or supremum for Y , denoted $\sup Y$, if x is an upper bound of Y and for all $y \in Y$ and all $z \in X, y \leq z$ implies $x \leq z$. The greatest upper bound or infimum for Y , denoted $\inf Y$, is defined dually.

Definition 5. A poset $\langle X, \leq \rangle$ is a lattice if, and only if, for all $x, y \in X$ both $\inf\{x, y\}$ and $\sup\{x, y\}$ exist in X . If $\langle X, \leq \rangle$ is a lattice, it is common to write $x \wedge y$, the meet of x and y , and $x \vee y$, the join of x and y , for $\inf\{x, y\}$ and $\sup\{x, y\}$, respectively. For $Y \subseteq X$, we use $\bigwedge Y$ and $\bigvee Y$ to denote the meet and join of all elements in Y , respectively.

Definition 6. For $a, b \in X$, we say that a is covered by b (or b covers a) if $a < b$ and $a \leq c < b \Rightarrow a = c$. The set of elements that are covered by b is denoted by b^- .

The classic example of a lattice is the power set of a set X ordered by inclusion, denoted $\langle \mathcal{P}(X), \subseteq \rangle$. Lattices are naturally represented by Hasse diagrams, in which nodes correspond to members of X and an edge exists between elements x and y if x covers y . FIG. S1A depicts the Hasse diagram for the lattice $\langle \mathcal{P}(X), \subseteq \rangle$ with $X = \{1, 2, 3\}$.

Definition 7. If $\langle X, \leq \rangle$ is a poset, $Y \subseteq X$ is a chain if for all $a, b \in Y$ either $a \leq b$ or $b \leq a$. Y is an antichain if $a \leq b$ only if $a = b$.

FIG. S1B shows examples of a chain and an antichain.

Definition 8. If there exists an element $\perp \in X$ with the property that $\perp \leq x$ for all $x \in X$, we call \perp the bottom element of X . The top element of X , denoted by \top , is defined dually.

Definition 9. For any $x \in X$, we define

$$\downarrow x = \{y \in X : y \leq x\} \text{ and } \dot{\downarrow} x = \{y \in X : y < x\}$$

where $\downarrow x$ and $\dot{\downarrow} x$ are called the down-set and strict down-set of x , respectively.

FIG. S1C illustrates the concepts of top and bottom elements, covering relations, and down-sets.

Appendix C: $\mathcal{A}(\mathbf{R})$ and the Redundancy Lattice

Formally, $\mathcal{A}(\mathbf{R})$ corresponds to the set of antichains on the lattice $\langle \mathcal{P}(\mathbf{R}), \subseteq \rangle$ (excluding the empty set). The cardinality of this set for $|\mathbf{R}| = n - 1$ is given by the $(n - 1)$ -th Dedekind number, which for $n = 2, 3, 4, \dots$ is 1, 4, 18, 166, 7579, \dots ([21], p. 273). The fact that $\langle \mathcal{A}(\mathbf{R}), \preceq \rangle$ forms a lattice, which we call the redundancy lattice, is proven in [46], where the corresponding lattice is denoted $\langle \mathcal{A}(X), \preceq' \rangle$ (see also [47]). As shown in [46], the meet (\wedge) and join (\vee) for this lattice are given by

$$\alpha \wedge \beta = \underline{\alpha \cup \beta} \quad (\text{A4})$$

and

$$\alpha \vee \beta = \underline{\uparrow \alpha \cap \uparrow \beta}. \quad (\text{A5})$$

Appendix D: Supporting Proofs

Theorem 1. $I(S = s; \mathbf{A})$ is nonnegative.

Proof.

$$I(S = s; \mathbf{A}) = D(p(\mathbf{a}|s) \parallel p(\mathbf{a})) \geq 0$$

where D is the Kullback-Leibler distance and the last step follows from the information inequality ([15], p. 26). \square

Lemma 1. $I(S = s; \mathbf{A})$ increases monotonically on the lattice $\langle \mathcal{P}(\mathbf{R}), \subseteq \rangle$.

Proof. Consider \mathbf{A}, \mathbf{B} with $\mathbf{A} \subset \mathbf{B} \subseteq \mathbf{R}$. Let $\mathbf{C} = \mathbf{B} \setminus \mathbf{A} \neq \emptyset$. Then we have

$$\begin{aligned}
& I(S = s; \mathbf{B}) - I(S = s; \mathbf{A}) \\
&= \sum_{\mathbf{b}} p(\mathbf{b}|s) \log \frac{p(s, \mathbf{b})}{p(s)p(\mathbf{b})} - \sum_{\mathbf{a}} p(\mathbf{a}|s) \log \frac{p(s, \mathbf{a})}{p(s)p(\mathbf{a})} \\
&= \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{a}, \mathbf{c})}{p(s)p(\mathbf{a}, \mathbf{c})} - \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{a})}{p(s)p(\mathbf{a})} \\
&= \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{c}|\mathbf{a})}{p(s|\mathbf{a})p(\mathbf{c}|\mathbf{a})} \\
&= \sum_{\mathbf{a}} p(\mathbf{a}) \sum_{\mathbf{c}} p(\mathbf{c}|\mathbf{a}, s) \log \frac{p(\mathbf{c}|\mathbf{a}, s)}{p(\mathbf{c}|\mathbf{a})} \\
&= \sum_{\mathbf{a}} p(\mathbf{a}) D(p(\mathbf{c}|\mathbf{a}, s) \parallel p(\mathbf{c}|\mathbf{a})) \geq 0.
\end{aligned}$$

□

Theorem 2. I_{\min} increases monotonically on the lattice $\langle \mathcal{A}(\mathbf{R}), \preceq \rangle$.

Proof. We proceed by contradiction. Assume there exists $\alpha, \beta \in \mathcal{A}(\mathbf{R})$ with $\alpha \prec \beta$ and $I_{\min}(S; \beta) < I_{\min}(S; \alpha)$. Then, from Eq. (3), there must exist $\mathbf{B} \in \beta$ such that $I(S = s; \mathbf{B}) < I(S = s; \mathbf{A})$ for some outcome $s \in S$ and for all $\mathbf{A} \in \alpha$. Thus, from Lemma 1, there does not exist $\mathbf{A} \in \alpha$ such that $\mathbf{A} \subseteq \mathbf{B}$. However, since $\alpha \prec \beta$ by assumption, there exists $\mathbf{A} \in \alpha$ such that $\mathbf{A} \subseteq \mathbf{B}$. □

Theorem 3. $\Pi_{\mathbf{R}}$ can be stated in closed form as

$$\Pi_{\mathbf{R}}(S; \alpha) = I_{\min}(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathbf{B} \subseteq \alpha^- \\ |\mathbf{B}|=k}} I_{\min}(S; \bigwedge \mathbf{B}). \quad (\text{A6})$$

Proof. For $\mathcal{B} \subseteq \mathcal{A}(\mathbf{R})$, define the set-additive function f as

$$f(\mathcal{B}) = \sum_{\beta \in \mathcal{B}} \Pi_{\mathbf{R}}(S; \beta).$$

From Eq. (6), it follows that $I_{\min}(S; \alpha) = f(\downarrow \alpha)$ and

$$\begin{aligned}
\Pi_{\mathbf{R}}(S; \alpha) &= f(\downarrow \alpha) - f(\downarrow \alpha) \\
&= f(\downarrow \alpha) - f\left(\bigcup_{\beta \in \alpha^-} \downarrow \beta\right).
\end{aligned}$$

Applying the principle of inclusion-exclusion ([20], p. 64), we have

$$= f(\downarrow \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathbf{B} \subseteq \alpha^- \\ |\mathbf{B}|=k}} f\left(\bigcap_{\gamma \in \mathbf{B}} \downarrow \gamma\right)$$

and it is a basic result of lattice theory that for any lattice L and $A \subseteq L$, $\bigcap_{a \in A} \downarrow a = \downarrow (\bigwedge A)$ ([44], p. 57), so we have

$$\begin{aligned}
&= f(\downarrow \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathbf{B} \subseteq \alpha^- \\ |\mathbf{B}|=k}} f(\downarrow (\bigwedge \mathbf{B})) \\
&= I_{\min}(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathbf{B} \subseteq \alpha^- \\ |\mathbf{B}|=k}} I_{\min}(S; \bigwedge \mathbf{B}).
\end{aligned}$$

□

Lemma 2 (Maximum-minimums identity). *Let A be a set of numbers. The maximum-minimums identity states that*

$$\max A = \sum_{k=1}^{|A|} (-1)^{k-1} \sum_{\substack{B \subseteq A \\ |B|=k}} \min B$$

or conversely,

$$\min A = \sum_{k=1}^{|A|} (-1)^{k-1} \sum_{\substack{B \subseteq A \\ |B|=k}} \max B.$$

Proof. It is proven in a number of introductory texts, e.g. [48]. □

Theorem 4. $\Pi_{\mathbf{R}}$ can be stated in closed form as

$$\Pi_{\mathbf{R}}(S; \alpha) = I_{\min}(S; \alpha) - \sum_s p(s) \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}). \quad (\text{A7})$$

Proof.

Combining Eqs. (A6) and (3) yields

$$\begin{aligned} \Pi_{\mathbf{R}}(S; \alpha) &= I_{\min}(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} \sum_s p(s) \min_{\mathbf{B} \in \bigwedge \mathcal{B}} I(S = s; \mathbf{B}) \\ &= I_{\min}(S; \alpha) - \sum_s p(s) \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} \min_{\mathbf{B} \in \bigwedge \mathcal{B}} I(S = s; \mathbf{B}) \end{aligned}$$

and by Lemma 1 and Eq. (A4),

$$= I_{\min}(S; \alpha) - \sum_s p(s) \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} \min_{\beta \in \mathcal{B}} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}).$$

Then, applying Lemma 2 we have

$$= I_{\min}(S; \alpha) - \sum_s p(s) \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}).$$

□

Theorem 5. $\Pi_{\mathbf{R}}$ is nonnegative.

Proof. If $\alpha = \perp$, $\Pi_{\mathbf{R}}(S; \alpha) = I_{\min}(S; \alpha)$ and $\Pi_{\mathbf{R}}(S; \alpha) \geq 0$ follows from the nonnegativity of I_{\min} . To prove it for $\alpha \neq \perp$, we proceed by contradiction. Assume there exists $\alpha \in \mathcal{A}(\mathbf{R}) \setminus \{\perp\}$ such that $\Pi_{\mathbf{R}}(S; \alpha) < 0$. Applying Eq. (3) to Theorem 4 and combining summations yields

$$\Pi_{\mathbf{R}}(S; \alpha) = \sum_s p(s) \{ \min_{\mathbf{A} \in \alpha} I(S = s; \mathbf{A}) - \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}) \}.$$

From this equation, it is clear that there must exist $\beta \in \alpha^-$ such that for all $\mathbf{B} \in \beta$, $I(S = s; \mathbf{A}) < I(S = s; \mathbf{B})$ for some outcome $s \in S$ and some $\mathbf{A} \in \alpha$. Thus, from Lemma 1, there does not exist $\mathbf{B} \in \beta$ such that $\mathbf{B} \subseteq \mathbf{A}$. However, since $\beta \prec \alpha$ by definition, there exists $\mathbf{B} \in \beta$ such that $\mathbf{B} \subseteq \mathbf{A}$. □

Appendix E: Supplementary Figures

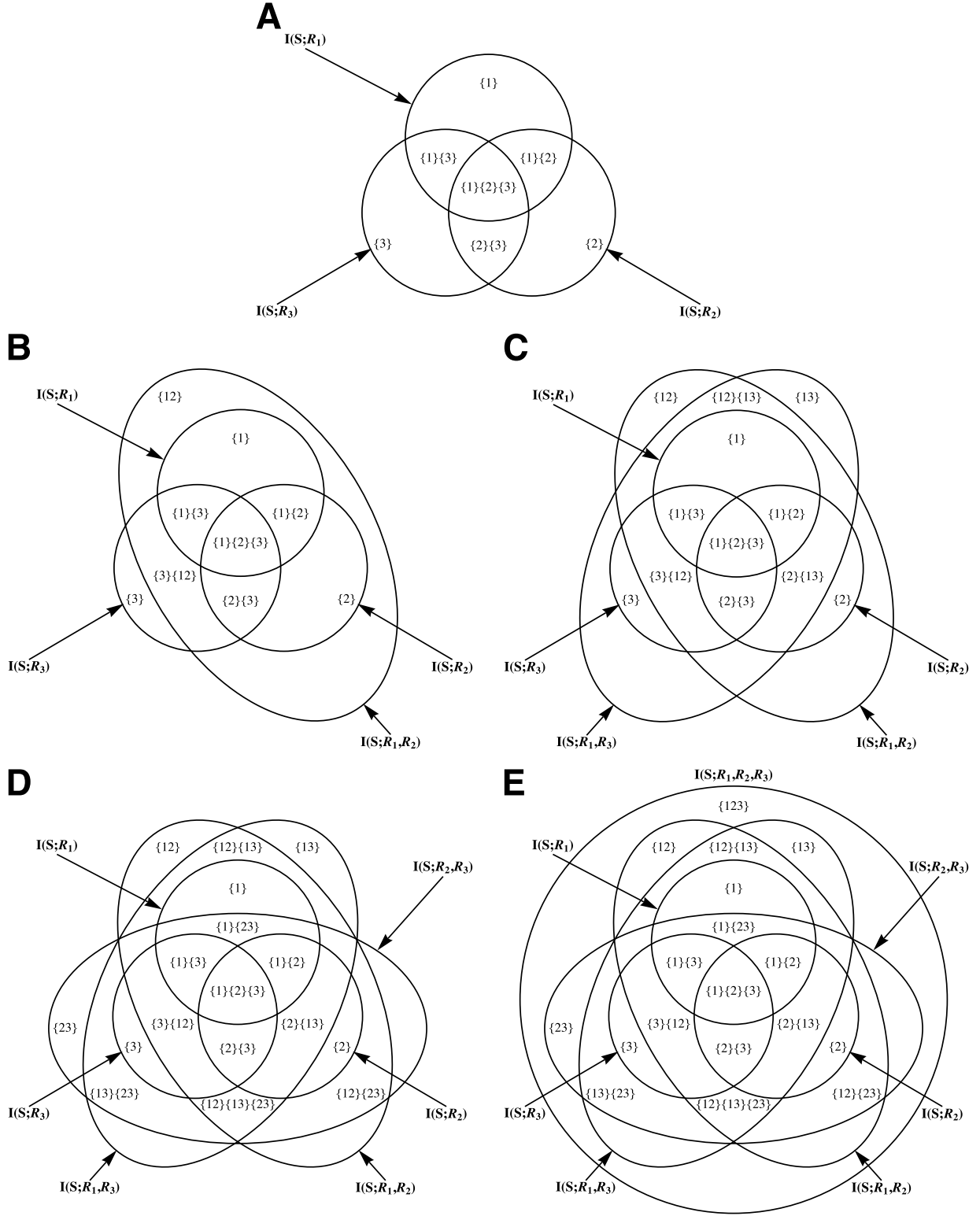


FIG. S2: Constructing a PI-diagram for 4 variables. (A) For each element $R_i \in \mathbf{R}$ there is a region corresponding to $I(S; R_i)$. (B-E) For each subset \mathbf{A} of \mathbf{R} with two or more elements, $I(S; \mathbf{A})$ is depicted as a region containing $I(S; A)$ for all $A \in \mathbf{A}$ but not coextensive with $\bigcup_{A \in \mathbf{A}} I(S; A)$. Regions of the diagram intersect generically, representing all possibilities for redundancy.

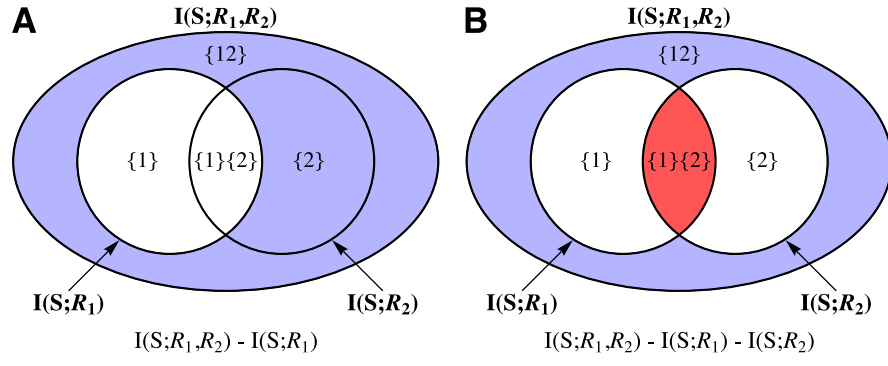


FIG. S3: Computing the PI-decomposition for 3-variable interaction information. (A-B) Term-by-term calculation of $I(S; R_1; R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2)$. Blue and red regions represent PI-terms that are added and subtracted, respectively.

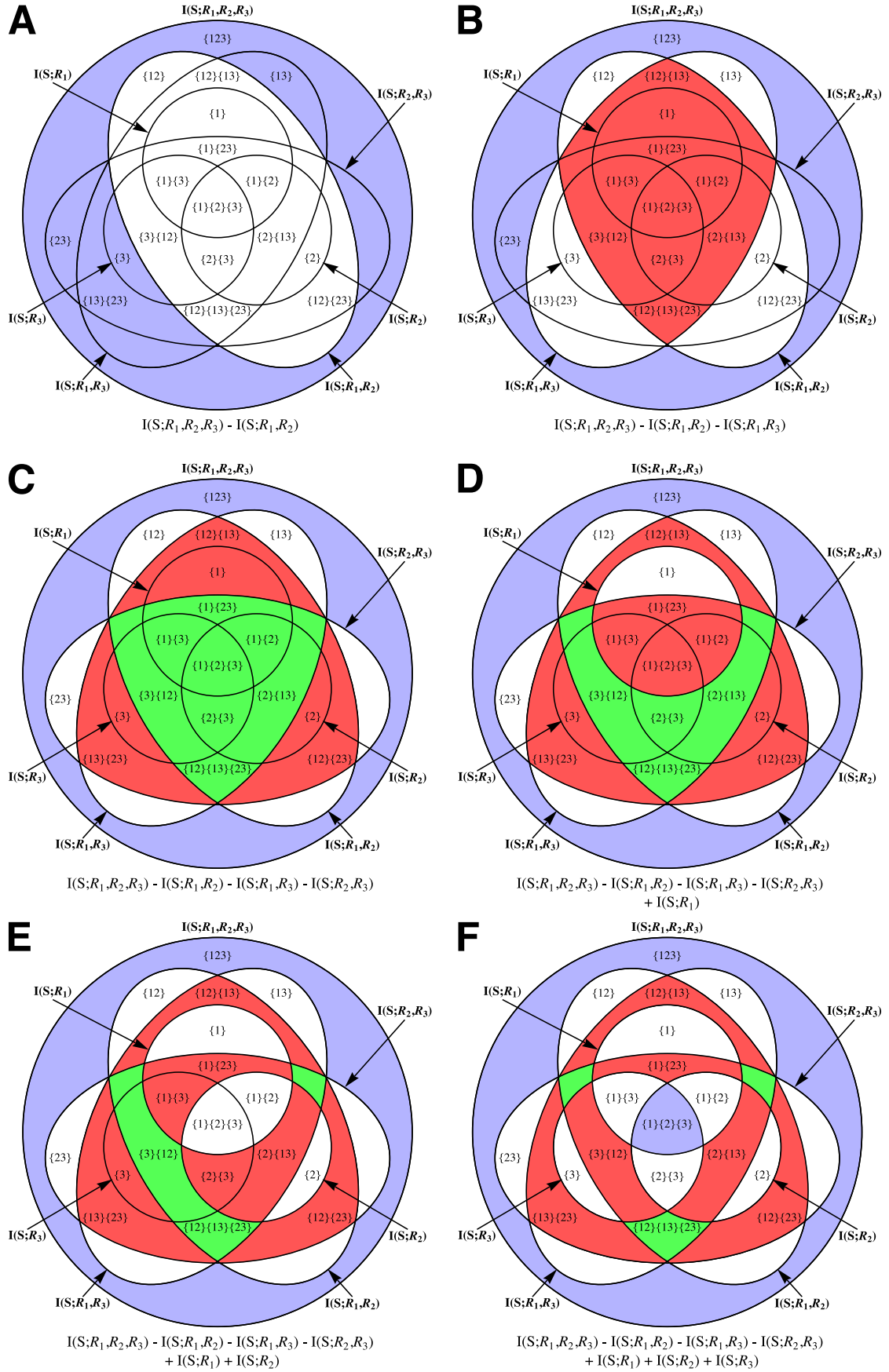


FIG. S4: Computing the PI-decomposition for 4-variable interaction information. (A-F) Term-by-term calculation of $I(S; R_1; R_2; R_3) = I(S; R_1, R_2, R_3) - I(S; R_1, R_2) - I(S; R_1, R_3) - I(S; R_2, R_3) + I(S; R_1) + I(S; R_2) + I(S; R_3)$. Blue and red regions represent PI-terms that are added and subtracted, respectively. Green regions represent PI-terms that are subtracted twice.